Enhancing Human Trajectory Prediction with Reinforcement Learning from Quantified Human Preferences

Chenyou Fan[1][0000−0002−9835−8507], Kehui Tan[1][0009−0001−5241−9028], Yanzhao Chen[1], Tianqi Pang[1], Haiqi Jiang[1], and Junjie Hu[2✉][0000−0002−1911−4361]

[1] South China Normal University, Guangdong, China
fanchenyou@scnu.edu.cn
[2] The Chinese University of HongKong, Shenzhen, China
hujunjie@cuhk.edu.cn

Abstract. We improve human trajectory prediction by introducing Reinforcement Learning from Human Feedback (RLHF) and Rejection Sampling techniques. To quantify human preferences, we parameterize and pre-train a diffusion backbone that models realistic human behaviors in the latent space. We then derive the diffusion score based on the latent trajectory feature, indicating the alignment between predicted trajectories and human decisions. By using the diffusion score as a reward, we refine the prediction model to generate trajectories preferred by humans. We further utilize rejection sampling to select the highest-scored trajectories to enhance the training. We validate our approach through various numerical experiments, human evaluations, and visualizations, showcasing a 15% reduction in positional deviation and a 20% increase in alignment with human preferences. Our proposed diffusion score can achieve a 67% Top-5 hit rate in retrieving the best candidate path with the least deviations from true human trajectories, thereby being capable of guiding realistic decision-making.

1 Introduction

The human trajectory prediction (HTP) task involves accurately forecasting the future behaviors of human drivers and pedestrians. Considerable progress has been made in developing advanced deep learning-based prediction models, including the utilization of Graph Convolutional Networks [20], Transformers [6], and recent Diffusion models [8].

Existing approaches often train HTP models by supervised regression against ground truth trajectories. However, this learning objective naturally limits it to evaluating how well the model aligns with overall human decision-making processes. Currently, there is no standardized way to quantify this alignment, making it challenging to design effective learning schemes that can better match human behavior. The above research challenges lead us to raise two questions: 1) *how to properly quantify human preference in a multi-human context with a*

parameterized model, and 2) *how to integrate this score into supervised trajectory modeling to enhance the learning process*.

To address the **first key question** of designing a robust scoring function to quantify human preference in Human Trajectory Prediction (HTP) task automatically, we design a robust scoring function that parameterizes human decisions as trainable models that can map predictions into a latent space where superior predictions closely align with ground truths. To achieve this, we leverage the advanced generative diffusion model [10] and quantify the alignment of predictions with true human decisions.

We diffuse the predictions in latent space and measure its distance from a random noise vector as residual. We normalize this residual by dividing it by the ground-truth residual, and then subtracting this ratio from one. The designed *diffusion score* quantifies how well the predicted trajectories align with actual human decisions: a higher score indicates a higher likelihood of accurate predictions.

To address the **second key question** of *how to integrate this score into supervised trajectory modeling to enhance the learning process*, We propose a comprehensive pipeline for refining the trajectory prediction diffusion model based on the Reinforcement Learning from Human Preference (RLHF) methodology. We design the reward feedback from the predicted trajectories based on their diffusion scores as a surrogate for human preferences. We guide RLHF procedures using Proximal Policy Optimization (PPO) [25] for fine-tuning the prediction model parameters.

Existing approaches typically generate multiple stochastic samples as candidate predictions. However, we observe that some samples fail to align with true human decisions, exhibiting significant deviations from the expected outcomes. This motivates us to develop RLHF-ReS, which applies rejection sampling to filter out deviated negative samples and performs policy gradient updates only on top-scoring trajectories, refining best samples by 10% in final deviations without sacrificing diversity.

We conduct thorough numerical experiments and human evaluations to validate our methods which achieved substantial gains in trajectory prediction accuracy, with average displacement improved by 3% and final displacement by 15%. Ablation studies confirm that the diffusion score serves as a superior evaluation metric and also provides a reliable ranking that can be used independently as an unsupervised metric. This property makes it well-suited for real-world decision-making scenarios.

In summary, the main contributions of our work include:

- **Quantified human preference.** We quantify human preference by measuring trajectories in latent space and developing a diffusion-based rewarding function.
- **RLHF with rejection sampling.** Our RLHF framework designs to incentivize models to improve trajectory generation through our targeted rewards.
- **Numerical results.** Our RLHF-ReS method achieves a 15% boost in positional accuracy and a 20% increase in alignment with human preference by survey statistics.

– **Feasible realistic decision-making.** Our proposed diffusion score achieves a Top-5 67% recall rate of identifying the best prediction sample.
– **In-context learning.** We perform continual adaptation of model weights to longer context windows to obtain an additional 6-9% improvement.

2 Related Work

Trajectory Prediction has been extensively studied due to its importance in autonomous driving and behavior prediction. Classical predicting methods learn patterns from historical path and predict the future, with Time-Series generative models as RNNs [1, 9, 13, 32, 35], GANs [7, 9, 24] and GNNs [20, 26, 33], VAE [34], and Transformers [5, 27, 36].

Gu *et al.* [8] proposed Motion indeterminacy diffusion (MID) to utilize diffusion models [28, 29] for generating trajectory predictions. Li *et al.* [15] proposed a bidirectional diffusion framework. Mao *et al.* [19] proposed Leapfrog Diffusion Model (LED) which learns to emit coarse future trajectory samples directly, then applies a few more denoising steps to save major time of trajectory generation.

Unlike existing works, we propose a systematic framework that enables fine-tuning the generation process through quantified human feedback, rather than relying solely on supervised regression on the positions.

Reinforcement learning from human feedback (RLHF) was widely applied in robot simulation [3] and game playing [11]. RLHF has recently been effectively adopted in fine-tuning LLMs such as GPTs to reduce risks, reject harmful contents, and improve human readability outputs [2, 4, 12, 21, 30]. Recent studies [17, 31] showed **Rejection Sampling** (ReS) can further enhance the RLHF through fine-tuning from the highest human-rewarded samples, such as in LLAMA [31] training.

Unlike traditional RLHF and ReS methods, we distill human feedback from their actual behavior rather than subjective ratings, quantifying their decisions to enhance stochastic sampling while maintaining diversity and accuracy.

3 Diffusion-Based Prediction Preliminaries

Human Trajectory Prediction (HTP) task. We observe P participating humans about their 2-D historical trajectories $\boldsymbol{X} = \{X_1, X_2, ..., X_{T_h}\} \in \mathbb{R}^{P \times T_h \times 2}$ over the past T_h steps. The HTP aims to estimate the future trajectories $\hat{\boldsymbol{Y}} = \{\hat{Y}_1, \hat{Y}_2, ..., \hat{Y}_{T_f}\} \in \mathbb{R}^{P \times T_f \times 2}$ for the consecutive future T_f steps with minimized displacements from the ground-truth coordinates $\boldsymbol{Y}$. Existing approaches commonly predict K stochastic samples (i.e., 20) for each participant for enclosing possible future paths.

We briefly summarize existing diffusion-based HTP models, including MID [8] and LED [19]. We abstract their architectures as two parts: the positional distribution generator G and the future trajectory diffusion model D.

The positional generator G encodes the social patterns from the observed history $\boldsymbol{X}$. Then G estimates each human's stepwise future trajectory distribution,

including global mean $\boldsymbol{\mu} \in \mathcal{R}^{P \times T_f \times 2}$, K stochastic sample shifts $\boldsymbol{S} \in \mathcal{R}^{K \times P \times T_f \times 2}$ and variances $\boldsymbol{\sigma} \in \mathcal{R}^{K \times P}$, and re-parameterize future trajectories as:

$$\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{S} \leftarrow G(\boldsymbol{X}) \ , \ \hat{\boldsymbol{Y}}^{\tau} = \boldsymbol{\mu} + \boldsymbol{\sigma} \cdot \boldsymbol{S} \ . \tag{1}$$

The diffusion model D aims to denoise the coarse predictions $\hat{\boldsymbol{Y}}^{\tau}$ with learned priors of human decisions. We denote its forward process as D^f and reverse process as D^r.

In forward process, D^f diffuses raw trajectories into Gaussian noises with fixed schedules. The training of D follows DDPM objective [10] which compares the diffused trajectories with random noises:

$$\mathcal{L}^D = \mathbb{E}_{X,Y} ||D^f(\boldsymbol{X}, \boldsymbol{Y}) - \boldsymbol{Z}||_2^2 \ , \tag{2}$$

in which $D^f(\boldsymbol{X}, \boldsymbol{Y})$ stands for the stepwise process of $\boldsymbol{Y}_{t+1} = \sqrt{1-\beta_t}\boldsymbol{Y}_t + \sqrt{\beta_t}\boldsymbol{\epsilon}_t$ that gradually adds Gaussian noises $\boldsymbol{\epsilon}_t \sim \mathcal{N}(0, 1)$ to the trajectories with a fixed β-schedules.

In *reverse* process, D^r sequentially denoises the coarse predictions $\hat{\boldsymbol{Y}}^{\tau}$ with τ denoising steps and produces refined prediction $\hat{\boldsymbol{Y}}$ as (from right to left):

$$\hat{\boldsymbol{Y}} \Leftrightarrow \hat{\boldsymbol{Y}}^0 \leftarrow \hat{\boldsymbol{Y}}^1 \leftarrow \cdots \hat{\boldsymbol{Y}}^{\tau-1} \leftarrow D^r(\boldsymbol{X}, \hat{\boldsymbol{Y}}^{\tau}) \ . \tag{3}$$

In training, the generator G, which produces raw $\hat{\boldsymbol{Y}}^{\tau}$, can be optimized through D by regressing the denoised $\hat{\boldsymbol{Y}}$ in Eq. (3) with ground-truth future $\boldsymbol{Y}$:

$$\mathcal{L}^{reg} = \mathbb{E}_{K,P,T_f} ||\mathbf{Y} - \hat{\mathbf{Y}}||_2^2 \ . \tag{4}$$

4 Trajectory Score (TS) of human preference

We propose a diffusion-based trajectory scoring function as a robust measure to quantify the alignment between predictions and human preferences. Using this score, we introduce two novel evaluation metrics that can serve as effective decision-making signals, even without true human preferences.

We define the *diffusion residual* for prediction $\hat{\boldsymbol{Y}}$ based on history $\boldsymbol{X}$ for each future step $t \in \{1, ..., T_f\}$, as the deviation from noise $\boldsymbol{Z} \sim \mathcal{N}(0, 1)$ as follows:

$$\boldsymbol{d}_p(\hat{\boldsymbol{Y}}; \boldsymbol{X}) = ||D^f(\boldsymbol{X}, \hat{\boldsymbol{Y}}) - \boldsymbol{Z}||_2^2 \in \mathcal{R}^{K \times P \times T_f} \ . \tag{5}$$

Where $D^f(\boldsymbol{X}, \hat{\boldsymbol{Y}})$ diffuses future trajectories into a latent space that embeds realistic human trajectories. A smaller residual $\boldsymbol{d}_p$ indicates a more realistic prediction $\hat{\boldsymbol{Y}}$ which aligns better with human decisions. Likewise, we calculate the residual for the ground-truth trajectory $\boldsymbol{Y}$ as an anchor denoted by $\boldsymbol{d}_g(\boldsymbol{Y}; \boldsymbol{X})$.

We define the *trajectory score* $\boldsymbol{r}$ (the larger the better) as subtracting the residual $\boldsymbol{d}_p$ normalized by $\boldsymbol{d}_g$ from one as:

$$\boldsymbol{r}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}; \boldsymbol{X}) = \left(1 - \frac{\boldsymbol{d}_p(\hat{\boldsymbol{Y}}; \boldsymbol{X})}{\boldsymbol{d}_g(\boldsymbol{Y}; \boldsymbol{X})}\right) \in \mathcal{R}^{K \times P \times T_f} \ . \tag{6}$$

A smaller residual $\boldsymbol{d}_p$ results in a larger score $\boldsymbol{r}$, indicating a more favorable prediction. This designed score provides a robust signal for both evaluation and model refinement through RLHF, as detailed in Sec. 5.

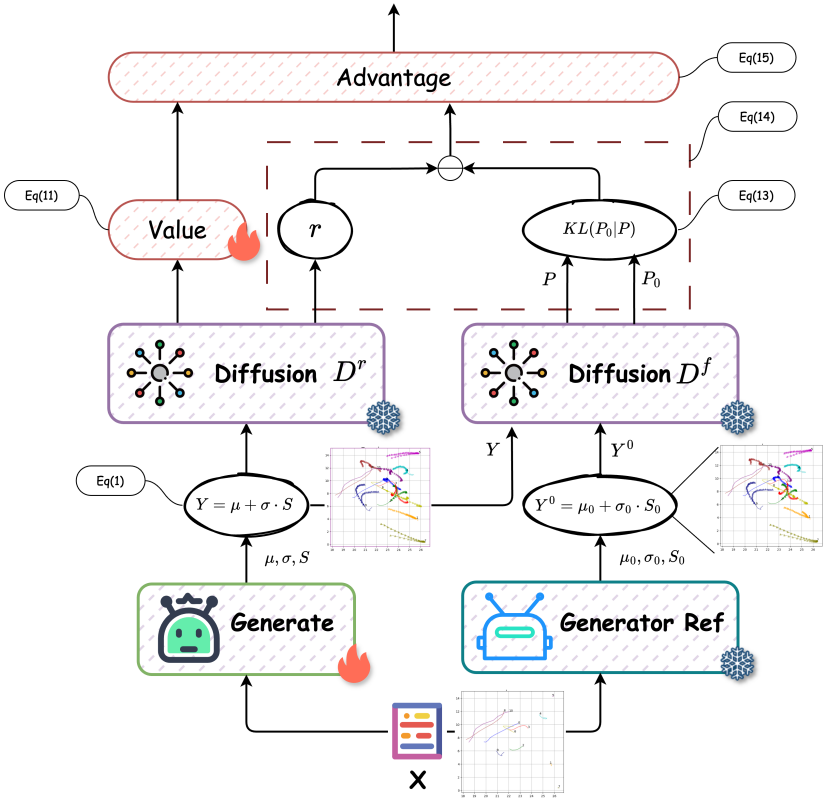

Fig. 1: Our diffusion-score-based RLHF framework. We freeze diffusion model D and reference generator model. We fine-tune the generator G during RLHF.

4.1 Two novel evaluation metrics

Conventional studies [1, 9, 23] adopt the Average Displacement Error (ADE) and Final Displacement Error (FDE), to evaluate prediction accuracy through $L2$-distance between predictions and ground-truth positions, averaged over all future steps (ADE) or only at the final step (FDE). However, these metrics have limitations in capturing true human preferences.

Instead, we propose the Average and Final *Trajectory diffusion Score*, abbreviated as **ATS** and **FTS**, which are the higher the better. These two novel metrics enable explicit and effective assessment of predicted multi-person trajectories based on their overall alignment with human preferences.

For prediction $\hat{\boldsymbol{Y}}$, the ATS for k-th trajectory sample is the trajectory score $\boldsymbol{r}$ averaged over all T_f future steps and all P humans in the scene such as:

$$ATS(k) = \frac{1}{P \cdot T_f} \sum_{p=1}^{P} \sum_{t=1}^{T_f} \boldsymbol{r}(k, p, t) \ , \tag{7}$$

while the FTS is the score at the final step T_f such as:

$$FTS(k) = \frac{1}{P}\sum_{p=1}^{P} \boldsymbol{r}(k, p, T_f) \ . \tag{8}$$

In Lemma 1, we prove that both ATS and FTS can be properly utilized to make decisions in real-time scenarios.

Lemma 1. *The overall ranking among all stochastic samples, as measured by ATS (or FTS), is independent of the ground truth. As such, we can select the best sample without the actual human decisions in advance.*

Proof. By definition of ATS as in Eq.(7), we will show that for any pair of two distinct samples among K stochastic predictions, such as the u-th and v-th sample, it does not rely on ground-truth information to determine if $\text{ATS}(u) > \text{ATS}(v)$, or equivalently if $\sum_{p,T_f} \boldsymbol{r}(u) > \sum_{p,T_f} \boldsymbol{r}(v)$. Given Eq.(6), we have

$$\boldsymbol{r}(u) = 1 - \frac{\sum_{p,T_f} \boldsymbol{d}_p(u)}{\sum_{p,T_f} \boldsymbol{d}_g} \ , \quad \boldsymbol{r}(v) = 1 - \frac{\sum_{p,T_f} \boldsymbol{d}_p(v)}{\sum_{p,T_f} \boldsymbol{d}_g} \ . \tag{9}$$

Also, we know that two samples of a same scene share the ground-truth $\boldsymbol{Y}$, with its residual $\sum_{p,T_f} \boldsymbol{d}_g > 0$ as in Eq.(5). Therefore, the $\sum_{p,T_f} \boldsymbol{d}_p(u) < \sum_{p,T_f} \boldsymbol{d}_p(v)$ immediately implies $\sum_{p,T_f} \boldsymbol{r}(u) > \sum_{p,T_f} \boldsymbol{r}(v)$. □

Lemma 1 demonstrates that it is possible to obtain a ranking over all stochastic samples by comparing their diffusion residuals $\boldsymbol{d}_p$, without referencing ground-truth information. This enables a straightforward selection of the best sample for online decision-making in autonomous driving scenarios.

5 RLHF for trajectory fine-tuning

The diffusion score leads to design an RLHF framework that enhances the generation of trajectories with scores as feedback. This refinement process fine-tunes the generator G to embed quantified human preferences given scene context. We decompose the framework as four main components as shown in Fig. 1.

Trajectory distribution generator G_θ. G_θ is pre-trained in a supervised manner. G_θ is structured with multiple transformer layers to capture social interactions and feed-forward layers to output future distribution parameters. During RLHF, G_θ serves as a policy model that self-refines to generate improved trajectory samples that adhere to human decisions.

We also instantiate a reference generator G_0 which will be kept frozen during the entire fine-tuning process. G_0 functions to prevent G_θ from deviating too much. The G_θ is shown in blue in Fig.1; the G_0 is shown in light blue.

Critic model. The critic's value function, V_ψ, serves to evaluate the generated predictions. Inspired by AlphaGo, we devise the critic to share the same social-encoder backbone as the policy network, G_θ, while having its own trainable value-head, V_ψ (a 2-layer MLP). The trainable V_ψ is illustrated in red in Fig.1.

This value function, $V_\psi(\boldsymbol{X}, \hat{\boldsymbol{Y}})$, evaluates the state-value of each intermediate step of a trajectory $\hat{\boldsymbol{Y}}$ given the history $\boldsymbol{X}$.

Diffuser D serves a two-fold role. The forward process D^f provides a diffusing score in Eq.(6) as a reward, as shown in light purple Fig. 1. The reverse process D^r denoises and finalizes the generator's predictions following Eq.(3), as shown in dark purple. D is pre-trained by Eq.(2), which will be kept frozen during RLHF.

5.1 Step-A: Predict future state values

In each iteration, we randomly sample a scene comprising both the historical trajectories $\boldsymbol{X}$ of P humans and their corresponding ground-truth future paths $\boldsymbol{Y}$. The pre-trained generator G_θ estimates their future trajectories set $\hat{\boldsymbol{Y}}^\tau$, as Eq.(1), which consist of K stochastic samples, each representing trajectories of all P agents over T_f future steps.

We apply τ denoising steps to obtain refined predictions, as Eq.(3), and denote the above process as:

$$\hat{\boldsymbol{Y}} \leftarrow D^r(\boldsymbol{X}, \hat{\boldsymbol{Y}}^\tau) \ \in \mathbb{R}^{K\times P\times T_f\times 2} \ . \tag{10}$$

We utilize the value-head V_ψ to estimate the state value $\boldsymbol{V}$ of the predicted positions at all future steps, such as:

$$\boldsymbol{V} \leftarrow V_\psi(\hat{\boldsymbol{Y}}) \in \mathbb{R}^{K\times P\times T_f} \ . \tag{11}$$

We also obtain the reference prediction $\hat{\boldsymbol{Y}}_0$ by generator G_0 and its reference state value $\boldsymbol{V}_0$ estimated in the same fashion. This is shown in Step-A of Fig.1.

5.2 Step-B: Estimate stepwise rewards

Given $\hat{\boldsymbol{Y}}$'s diffusion score $\boldsymbol{r}$ estimated by Eq.(6), we aim to estimate the stepwise prediction reward for policy update.

Following PPO [25], we apply a regularization to limit the KL-divergence between G_θ and the reference model G_0. Given the predicted distributions $\boldsymbol{\mu}, \boldsymbol{\sigma}$ emitted from G_θ (or G_0), the likelihood $\boldsymbol{p}_\theta \in \mathbb{R}^{K\times P\times T_f}$ (or $\boldsymbol{p}_0$) are:

$$\boldsymbol{p}_\theta \leftarrow \ \mathcal{N}(\hat{\boldsymbol{Y}}; \boldsymbol{\mu}, \boldsymbol{\sigma}) \ , \text{ i.e. } \ \boldsymbol{p}_\theta = \frac{1}{\boldsymbol{\sigma}\sqrt{2\pi}} e^{\frac{-\left(\hat{\boldsymbol{Y}}-\boldsymbol{\mu}\right)^2}{2\boldsymbol{\sigma}^2}} \ . \tag{12}$$

Thus, the policy KL-divergence, with coefficient β, is as:

$$\boldsymbol{div}(\boldsymbol{p}_0||\boldsymbol{p}_\theta) = \beta \cdot \boldsymbol{p}_0 \cdot [\log(\boldsymbol{p}_0) - \log(\boldsymbol{p}_\theta)] \ . \tag{13}$$

In summary, the combined stepwise prediction reward $\boldsymbol{R}(t)$ is the diffusion score penalized by the divergence as $\boldsymbol{R}(t) = \boldsymbol{r}(t) - \boldsymbol{div}(t) \ \ , \forall t \in [1, ..., T_f]$.

Following the classic policy gradient-based reinforcement learning, we calculate the *advantage* of the prediction at each step t as measurement.

Given the reward $\boldsymbol{R}$ and state-value $\boldsymbol{V}$ in Eq. (11), we calculate the *advantage* $\boldsymbol{A}$ as the difference of action-value and state-value at each step t as follows:

$$\boldsymbol{A}(t) = \underbrace{\boldsymbol{R}(t) + \gamma \boldsymbol{V}(t+1)}_{\text{action-value}} - \underbrace{\boldsymbol{V}(t)}_{\text{state-value}} \ , \tag{14}$$

where the action-value decomposes to current reward $\boldsymbol{R}(t)$ plus next-state value $\boldsymbol{V}(t+1)$ scaled by a discount rate γ.

The advantage $\boldsymbol{A}(t) \in \mathbb{R}^{K \times P}$ quantifies the quality of each stochastic trajectory sample k for each participant p. A positive advantage indicates a favorable prediction that aligns with human choices and is otherwise unfavorable.

5.3 Step-C: Optimize policy and value network

To finalize RLHF, we update both the policy generator G_θ and the value head V_ψ by maximizing the advantage of each generated trajectory.

We design the training objective of policy generator G_θ concerning advantage and likelihood as:

$$\mathcal{L}_\theta^{pg} = -\mathbb{E}[\min\{\frac{\boldsymbol{p}}{\boldsymbol{p}_0} \cdot \boldsymbol{A}\ , \text{clip}(\frac{\boldsymbol{p}}{\boldsymbol{p}_0}, 1-\epsilon, 1+\epsilon) \cdot \boldsymbol{A}\}]. \tag{15}$$

where the clip function constrains the magnitude of the policy update. $\mathcal{L}_\theta^{pg}$ has two primary objectives: boosting realism by scaling factor $\frac{\boldsymbol{p}}{\boldsymbol{p}_0}$, and improving human preference by advantage $\boldsymbol{A}$.

Next, we optimize the state value head V_ψ to accurately estimate the value of predictions. Given the estimated state values $\boldsymbol{V}$ produced by V_ψ, we aim to minimize the discrepancy between these predicted values and the actual prediction rewards $\boldsymbol{R}$. Formally, we design the value loss function as:

$$\mathcal{L}_\psi^{val} = \mathbb{E}[\max\{(\boldsymbol{V} - \boldsymbol{R})^2, (\hat{\boldsymbol{V}} - \boldsymbol{R})^2\}]\ , \tag{16}$$

where $\hat{\boldsymbol{V}} = \text{clip}(\boldsymbol{V}, \boldsymbol{V}_0 - \epsilon, \boldsymbol{V}_0 + \epsilon)$ constrains the update, and $\boldsymbol{V}_0$ is the value from reference prediction $\hat{\boldsymbol{Y}}_0$.

Finally, we compose the RLHF multi-task learning objective by integrating the policy gradient loss $\mathcal{L}_\theta^{pg}$ in Eq.(15), value-head loss $\mathcal{L}_\psi^{val}$ in Eq.(16), as well as the positional regression loss $\mathcal{L}_\theta^{reg}$ in Eq.(4), as:

$$\mathcal{L}_{\theta,\psi}^{rl} = \mathcal{L}_\theta^{reg} + \alpha_p \mathcal{L}_\theta^{pg} + \alpha_v \mathcal{L}_\psi^{val}\ . \tag{17}$$

We adopt an end-to-end model training fashion by minimizing the multi-task objective $\mathcal{L}^{rl}$ with standard SGD and determine optimal scaling factors α_p and α_v through grid search over validation set.

6 RLHF with Rejection Sampling

We are further motivated by an observation that the stochastic predictions $\hat{\boldsymbol{Y}}$ emitted by the generator G_θ are diverse but infeasible, exhibiting significant deviations from the expected outcomes. This motivated us to develop a novel rejection sampling method referred to as **RLHF-ReS**, which enables the filtering of highly deviated negative samples during the RLHF policy update process.

Let k^* be the index of the best prediction $\hat{\boldsymbol{Y}}^{k^*} \in \mathbb{R}^{P \times T_f}$ among all K stochastic samples. The k^*-sample receives the maximum collective diffusion score

$\boldsymbol{r}$, integrated over each future step for all participants. Concretely, given stepwise diffusion score $\boldsymbol{r}$ in Eq.(6), k^* is chosen as $k^* \leftarrow \arg\max_K \sum_{p=1}^{P} \sum_{t=1}^{T_f} \boldsymbol{r}_{k,p,t}$.

RLHF-ReS applies RLHF fine-tuning only with the best scene k^* while rejecting others. We therefore re-write the policy gradient loss $\mathcal{L}_\theta^{pg}$ in Eq.(15) as:

$$\mathcal{L}_\theta^{pg*} = -\mathbb{E}\left[\min\left\{ \frac{\boldsymbol{p}^*}{\boldsymbol{p}_0^*} \cdot \boldsymbol{A}^* \, , \mathrm{clip}\left(\frac{\boldsymbol{p}^*}{\boldsymbol{p}_0^*}, 1-\epsilon, 1+\epsilon\right) \cdot \boldsymbol{A}^* \right\} \right] ; \tag{18}$$

as well as the state-value loss $\mathcal{L}_\psi^{val}$ in Eq.(16) as:

$$\mathcal{L}_\psi^{val*} = \max\left\{ (\boldsymbol{V}^* - \boldsymbol{R}^*)^2, (\hat{\boldsymbol{V}}^* - \boldsymbol{R}^*)^2 \right\} , \tag{19}$$

where the superscript $*$ indicates selecting only the best sample among all stochastic samples. Finally, the joint training objective of RLHF-ReS is:

$$\mathcal{L}^{rl*} = \mathcal{L}_\theta^{reg} + \beta_p \mathcal{L}_\theta^{pg*} + \beta_v \mathcal{L}_\psi^{val*} \ . \tag{20}$$

By applying rejection sampling, we preserve the model's ability to generate diverse predictions while enhancing the alignment with human decisions.

7 Experiments

7.1 Datasets and Methods

The pedestrian trajectory datasets include the **ETH** [22] and **UCY** [14] which comprises a total of 5 subsets, as well as the **Stanford Drone Dataset (SDD)** [23] collected from campus streets from the sky. Following previous works [1, 18, 33], each data sequence of the above datasets consists of observed trajectories spanning 8 frames (3.2 seconds) and future trajectories of the next 12 frames (4.8 seconds).

The **NBA** [16] dataset is gathered from publicly available NBA games. Each sequence records the movements of the 10 players and the ball. Each data sequence consists of observed trajectories spanning 10 frames (2 seconds) and future trajectories of subsequent 20 frames (4 seconds).

We collect a **Robo** dataset which simulates a scenario where two robots are walking together, while another robot approaches them and adjusts its path to avoid collision. Each episode has a history of 8 frames (4 seconds) and a future of 12 frames (6 seconds).

Our methods include **RLHF** which integrates RLHF with diffusion scoring (Sec. 5) and **RLHF-ReS** which further utilizes rejection sampling to improve policy learning (Sec. 6). **The baseline methods** include Social-LSTM [1], GAN-based SocialGAN [9], GNN-based STGCNN [20], GroupNet [33], SGCN [26] and SGCN-SDA [5]. The diffusion-based models include MID [8] and LED [19].

Implementation details. The denoising process uses 100 steps with fast inference sampling every 3 steps, following the standard DDPM framework. For noise scheduling, we use a linear interpolation for β, ranging from 10^{-4} (initial) to $5 \cdot 10^{-2}$ (final).

Through grid search, the optimal scaling factors are $\alpha_p = 1.0$ and $\alpha_v = 5.0$ in RLHF objective Eq.(17), and $\beta_p = 0.3$ and $\beta_v = 1.5$ for RLHF-ReS in Eq.(20). The training process utilized GTX 3080Ti, running for 100 epochs over 12 hours on the ETH-UCY datasets and 24 hours on the NBA, SDD, and Robo datasets. The G follows LED [19] with a two-layer transformer-based social encoder with a hidden size 256.

7.2 Result analysis

Table 1: The minADE_{20} | minFDE_{20} (↓) results of trajectory prediction on the benchmark datasets.

	SGCN [26]	SDA [5]	GrpNet [33]	MID [8]	LED [19]	RLHF(ours)	RLHF-ReS
ETH	0.63/1.03	0.55/0.86	0.46/0.73	0.43/0.69	0.33/0.60	0.34/0.53	**0.31**/**0.48**
HOTEL	0.32/0.55	0.28/0.44	**0.15**/**0.25**	0.19/0.27	0.20/0.40	0.19/0.31	0.17/0.30
UNIV	0.37/0.70	0.37/0.69	**0.26**/0.49	**0.26**/**0.43**	0.31/0.62	0.30/0.56	0.29/0.48
ZARA1	0.29/0.53	0.27/0.46	0.21/0.39	0.23/0.39	**0.18**/0.31	0.19/**0.30**	**0.18**/**0.30**
ZARA2	0.25/0.45	0.24/0.37	0.17/0.33	0.21/0.32	0.16/0.28	**0.15**/**0.25**	**0.15**/0.26
NBA	1.09/1.96	0.92/1.21	1.13/1.69	0.96/1.27	0.80/1.11	0.79/1.02	**0.77**/**1.00**
Robo	0.43/0.70	0.41/0.66	0.43/0.68	0.29/0.40	0.22/0.35	0.22/0.36	**0.21**/**0.29**
AVG	0.48/0.85	0.43/0.67	0.40/0.65	0.37/0.54	0.31/0.52	0.31/0.48	**0.30**/**0.44**

Table 2: The maxATS_{20} and maxFTS_{20} (↑) on the benchmark datasets.

	SGCN [26]	SDA [5]	GrpNet [33]	MID [8]	LED [19]	RLHF(ours)	RLHF-ReS
ETH	0.16/-0.14	0.13/-0.15	0.21/0.02	0.39/0.12	0.50/0.18	0.50/0.19	**0.54**/**0.21**
HOTEL	0.05/-0.85	-0.02/-0.92	0.01/-0.87	0.05/-0.62	0.08/-0.66	0.14/-0.53	**0.19**/**-0.38**
UNIV	0.13/-0.74	0.12/-0.72	0.15/-0.39	0.27/-0.33	0.46/**-0.20**	**0.47**/**-0.20**	**0.47**/-0.21
ZARA1	-0.02/-0.66	-0.05/-0.65	1.05/-0.65	0.31/-0.39	0.32/-0.37	0.33/-0.36	**0.35**/**-0.34**
ZARA2	0.00/-0.96	0.02/-1.06	0.17/-0.93	0.24/-0.45	0.33/-0.41	**0.38**/**-0.38**	0.37/-0.40
NBA	-3.32/-4.72	-2.20/-3.69	-0.71/-2.38	0.09/-1.21	0.23/-0.61	0.23/**-0.60**	**0.24**/**-0.60**
Robo	-1.35/-3.25	-0.94/-2.49	-0.31/-1.81	0.02/-1.07	0.07/-0.59	0.15/-0.39	**0.19**/**-0.31**
AVG	-0.62/-1.62	-0.42/-1.38	0.08/-1.00	0.20/-0.56	0.28/-0.38	0.31/-0.32	**0.34**/**-0.29**

We show ADE and FDE(↓) in Table 1. Following previous works, we report the minimum ADE and FDE over $K = 20$ stochastic samples. The **best** values are presented in bold, while the second best values are underlined. We observe:

1) Our RLHF-ReS achieves the lowest ADE and FDE among all methods, reducing previous state-of-the-art LED by 3.2% in ADE (0.30 vs. 0.31) and 15.4% in FDE (0.44 vs. 0.52), as shown in last column of Table 1.

2) RLHF-ReS achieves a substantial decrease in FDE such as 9.9% in NBA and 17.1% in Robo, as rejection sampling tends to reduce high deviations.

3) RLHF performs second-best among all methods, which ties with LED in ADE but reduces FDE by 10.5%, thanks to the RLHF fine-tuning procedures.

4) We observe a similar trend on SDD, where our RLHF-ReS outperforms LED [19] by 4%/16% in ADE/FDE.

We show evaluations with our proposed metrics ATS and FTS (↑) in Table 2, computing the maximum ATS and FTS over $K = 20$ samples.

1) Our RLHF-ReS also achieves the highest ATS and FTS, with a 0.34 in ATS and -0.29 in FTS, leading the existing LED $0.28/-0.38$ by 21.4% in ATS and 21.1% in FTS. This proves its better alignment with human decisions.

2) RLHF-ReS outperforms the second-place RLHF by 9.7%/9.4%, proving the efficacy of rejection sampling in enhancing human preferences.

3) Notably, maxATS_{20} scores are predominantly positive, whereas maxFTS_{20} scores are negative. Due to the definition in Eq. (6), the max ATS among 20 samples is likely to exceed the ground-truth value, while the final step score (FTS) is lower due to accumulated deviations.

7.3 Visualizations and case studies

We provide visualizations of predicted trajectories in Fig. 2. We display the top-2 best predictions in ATS, shading them with weights proportional to the scores. We place realistic human decisions in the first column. In the ETH dataset (1st row, 1st col.), the upper-right person (orange) had minimal dynamics, captured by one stochastic prediction of our RLHF-ReS (2nd col.) model.

The NBA scene on the 2nd row exhibits dynamic trajectories. The ball (1st col., No.10, purple line) was passing through attacking player No.1 (orange), then crossing defending player 6 (green) and player 5 (pink), before being received by player 4 (light blue). Both RLHF-ReS (2nd col.) and RLHF (3rd col.) could properly reason about this order of players. The LED method (4th column) failed to accurately capture the interaction between players No.5 and No.4.

7.4 Human Evaluations

We conducted a human evaluation of generated trajectories through a five-point Likert scoring survey. Participants were presented with 12 scenes featuring predictions from four methods and true human decisions. Twenty human judges rated each method on a scale of 1-5, assessing the social appropriateness and kinetics of the predicted paths.

Table 3 shows our proposed RLHF-ReS method achieves the highest human-evaluated scores, with an average rating of 3.66. Notably, the ranking of human scores aligns well with our ATS and FTS scores, with RLHF-ReS leading the way, followed by RLHF, LED, and SGCN-SDA.

Our statistical analysis reveals a significant difference between RLHF-ReS and other methods: RLHF-ReS is significantly better than LED and SGCN-SDA

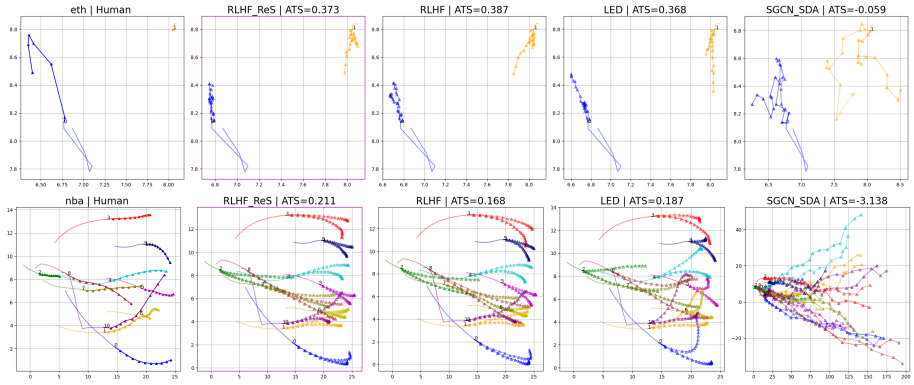

Fig. 2: Scenes from ETH and NBA. The first column displays human trajectories. The second to last columns show predictions from RLHF-ReS, RLHF, LED [19] and SGCN-SDA [5]. We display the top-2 most likely predictions for each person.

Table 3: Human evaluations, $*$ is significance level, ns is not significant.

Model	Human Score(↑)	ATS(↑)	FTS(↑)
RLHF-ReS	**3.66**	**0.44**	**-0.31**
RLHF	$\underline{3.58}^{ns}$	$\underline{0.40}^{**}$	$\underline{-0.33}^{ns}$
LED	3.28^{***}	$\underline{0.40}^{**}$	-0.41^{**}
SGCN-SDA	1.42^{***}	-0.57^{**}	-1.92^{**}

at a 1% level (indicated by ***) in human score, and also outperforms them at a 5% (indicated by **) level in ATS and FTS scores.

8 Ablation Studies

8.1 Hand-crafted reward. We employed a hand-crafted Social Distance Accuracy (SDA) reward, which clusters social groups and incentivizes close distances between those in the same group [5]. We observed that RLHF-SDA has a close ADE (0.32 vs. 0.30) but has a much larger FDE (0.53 vs. 0.44). This indicates the SDA reward leads to significant final-step deviations. Our approach mitigates this by projecting the global configuration into a latent space, providing more reliable reward signals.

8.2 Realistic decision making through ATS/FTS metrics. Lemma 1 shows that ATS (or FTS) yields a proper ranking of stochastic samples without ground truth knowledge, allowing us to investigate their confidence in retrieving optimal decisions. We take Top-1 and Top-5 ATS/FTS predictions and check if the target sample with minimum ADE/FDE is among them. We consider a successful recall if the target sample is retrieved. The overall recall rate measures ATS/FTS decision-making capability.

The Top-1 ATS recall rate (i.e., max ATS sample matches exactly min ADE) is 21.9%, increasing to 67.4% regarding Top-5 recall rate. Similarly, the Top-1 FTS recall rate is 18.4%, increasing to 57.4% for the Top-5 case. These results show that ATS and FTS metrics are suitable for real-time decision-making, providing reliable predictions for further planning.

8.4 In-context human decision prompting Inspired by prompting LLMs with examples to enhance the generation, we prompt our model with a longer context window and adapt it to recent scenes using RL rewards. During inference, we update the model weights using policy gradient after feeding it five consecutive segments (a 16-second window of 40 frames) and predict the next 12 frames.

We observed significant improvements in ADE and FDE across various datasets using RLHF-ReS with in-context learning. Specifically, in-context learning decreased ADE by 6.7% (0.28 vs. 0.30) and FDE by 9.1% (0.40 vs. 0.44). This outcome demonstrates the value of continuous model refinement to capture the scene context, a crucial consideration for applications like autonomous driving.

8.5 The number of participants. We evaluate RLHF on ZARA2 scenes with varying human counts (bins at 5, 10, 20). We find that RLHF consistently reduces ATS by 10-20% and FTS by 7-9%. The benefit slightly slides at more than 10 people due to scene complexity.

9 Conclusion

We devise a robust diffusion-based scoring function that approximates human preferences based on the diffusion residuals. With rejection sampling, we tailored the RLHF to refine the diffusion network. Our methods achieved an improvement of 5-20% in prediction precision and 10-20% enhancement in aligning the predictions with human preferences. Extensive ablation studies prove that the designed ADFs and FDFs metrics based on diffusion score are suitable for unsupervised decision making with high best trajectory recall rates.

Acknowledgments

This work is supported by the GuangDong Basic and Applied Basic Research Foundation (Project 2024A1515011650), Guangdong Natural Science Fund under Grant (Project 2024A1515010252), and the National Natural Science Foundation of China (Project 62106156).

References

1. Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S.: Social lstm: Human trajectory prediction in crowded spaces. In: CVPR (2016)
2. Bahdanau, D., Bengio, Y., et al.: An actor-critic algorithm for sequence prediction. arXiv preprint (2016)
3. Christiano, P.F., Leike, J., et al.: Deep reinforcement learning from human preferences. NIPS (2017)

4. Dong, R., Pang, W., Pan, C., Lu, H.y., Fan, C.: Storycrafter: Instance-aligned multi-character storytelling with diffusion policy learning. In: Proceedings of the 33rd ACM International Conference on Multimedia (2025)
5. Fan, C., Jiang, H., Huang, A., Hu, J.: Trajectory prediction with contrastive pre-training and social rank fine-tuning. ICONIP (2023)
6. Fang, L., Jiang, Q., Shi, J., Zhou, B.: Tpnet: Trajectory proposal network for motion prediction. In: CVPR (2020)
7. Gao, Q., Zhou, F., Zhang, K., Zhang, F., Trajcevski, G.: Adversarial human trajectory learning for trip recommendation. IEEE Transactions on Neural Networks and Learning Systems **34**(4), 1764–1776 (2021)
8. Gu, T., Chen, G., Li, J., Lin, C., Rao, Y., Zhou, J., Lu, J.: Stochastic trajectory prediction via motion indeterminacy diffusion. In: CVPR (2022)
9. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A.: Social gan: Socially acceptable trajectories with generative adversarial networks. In: CVPR (2018)
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NIPS (2020)
11. Ibarz, B., Leike, J., Pohlen, T., Irving, G., Legg, S., Amodei, D.: Reward learning from human preferences and demonstrations in atari. NIPS **31** (2018)
12. Kaufmann, T., Weng, P., Bengs, V., Hüllermeier, E.: A survey of reinforcement learning from human feedback. arXiv preprint arXiv:2312.14925 (2023)
13. Lee, N., Choi, W., et al.: Desire: Distant future prediction in dynamic scenes with interacting agents. In: CVPR (2017)
14. Lerner, A., Chrysanthou, Y., Lischinski, D.: Crowds by example. Computer Graphics Forum (2007)
15. Li, R., Li, C., Ren, D., Chen, G., Yuan, Y., Wang, G.: Bcdiff: Bidirectional consistent diffusion for instantaneous trajectory prediction. In: Advances in Neural Information Processing Systems. vol. 36 (2023)
16. Linou, K.: NBA Player Movements. https://github.com/linouk23/NBA-Player-Movements (2016), [Accessed 08-Nov-2023]
17. Liu, T., Zhao, Y., Joshi, R., Khalman, M., Saleh, M., Liu, P.J., Liu, J.: Statistical rejection sampling improves preference optimization. arXiv preprint arXiv:2309.06657 (2023)
18. Mangalam, K., Girase, H., Agarwal, S., Lee, K.H., Adeli, E., Malik, J., Gaidon, A.: It is not the journey but the destination: Endpoint conditioned trajectory prediction. In: ECCV (2020)
19. Mao, W., Xu, C., Zhu, Q., Chen, S., Wang, Y.: Leapfrog diffusion model for stochastic trajectory prediction. In: CVPR (2023)
20. Mohamed, A., Qian, K., Elhoseiny, M., Claudel, C.: Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In: CVPR (2020)
21. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. arXiv preprint arXiv:2203.02155 (2022)
22. Pellegrini, S., Ess, A., Schindler, K., van Gool, L.: You'll never walk alone: Modeling social behavior for multi-target tracking. In: ICCV (2009)
23. Robicquet, A., Sadeghian, A., Alahi, A., Savarese, S.: Learning social etiquette: Human trajectory understanding in crowded scenes. In: ECCV (2016)
24. Sadeghian, A., Kosaraju, V., Sadeghian, A., Hirose, N., Rezatofighi, H., Savarese, S.: Sophie: An attentive GAN for predicting paths compliant to social and physical constraints. In: CVPR (2019)
25. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)

26. Shi, L., Wang, L., Long, C., Zhou, S., Zhou, M., Niu, Z., Hua, G.: Sparse graph convolution network for pedestrian trajectory prediction. In: CVPR (2021)
27. Shi, L., Wang, L., Zhou, S., Hua, G.: Trajectory unified transformer for pedestrian trajectory prediction. In: International Conference on Computer Vision (2023)
28. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: ICML (2015)
29. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. NIPS (2019)
30. Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., Christiano, P.F.: Learning to summarize with human feedback. NIPS (2020)
31. Touvron, H., Lavril, T., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
32. Vemula, A., Muelling, K., Oh, J.: Social attention: Modeling attention in human crowds. In: ICRA (2018)
33. Xu, C., Li, M., Ni, Z., Zhang, Y., Chen, S.: Groupnet: Multiscale hypergraph neural networks for trajectory prediction with relational reasoning. In: CVPR (2022)
34. Xu, P., Hayet, J.B., Karamouzas, I.: Socialvae: Human trajectory prediction using timewise latents. In: ECCV (2022)
35. Yang, B., Yan, G., Wang, P., Chan, C.Y., Song, X., Chen, Y.: A novel graph-based trajectory predictor with pseudo-oracle. IEEE transactions on neural networks and learning systems **33**(12), 7064–7078 (2021)
36. Zhou, Z., Wang, J., Li, Y.H., Huang, Y.K.: Query-centric trajectory prediction. In: CVPR (2023)